Published in final edited form as:

J Mem Lang. 2014 August 1; 75: 181–198. doi:10.1016/j.jml.2014.06.003.

Test Framing Generates a Stability Bias for Predictions of Learning by Causing People to Discount their Learning Beliefs

Robert Ariel, Jarrod C. Hines, and Christopher Hertzog

Georgia Institute of Technology

Abstract

People estimate minimal changes in learning when making predictions of learning (POLs) for future study opportunities despite later showing increased performance and an awareness of that increase (Kornell & Bjork, 2009). This phenomenon is conceptualized as a stability bias in judgments about learning. We investigated the malleability of this effect, and whether it reflected people's underlying beliefs about learning. We manipulated prediction framing to emphasize the role of testing vs. studying on memory and directly measured beliefs about multi-trial study effects on learning by having participants construct predicted learning curves before and after the experiment. Mean POLs were more sensitive to the number of study-test opportunities when performance was framed in terms of study benefits rather than testing benefits and POLs reflected pre-existing beliefs about learning. The stability bias is partially due to framing and reflects discounted beliefs about learning benefits rather than inherent belief in the stability of performance.

Keywords

Metacognition; Metamemory; Monitoring; Stability Bias; Framing

Learners must make decisions about how to study material to effectively regulate their learning. These decisions may be informed by their expectations about how various factors will influence their later memory for material (Dunlosky & Ariel, 2011; Winne & Hadwin, 1998). Consider students studying for an upcoming exam. If they do not expect to achieve major gains in knowledge by studying material more than once, students may choose to stop studying before items are sufficiently mastered. Thus, it is important that learners have accurate knowledge of effective study techniques and of the conditions that influence their memory. However, knowledge alone does not always lead to effective strategy implementation (Borkowski, Carr, & Pressley, 1987).

© 2014 Elsevier Inc. All rights reserved

Please send correspondence to Robert Ariel, Georgia Institute of Psychology, School of Psychology, Atlanta, GA; robert.ariel@psych.gatech.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Footnotes

The relevant knowledge and beliefs that are in principle available to the learner can be distinguished from whether the learner accesses and uses them during learning. People do not always access relevant beliefs when predicting future learning. For instance, research on mnemonic strategy knowledge indicates that item-level judgments of learning (individuals' predictions of future memory for an item they just studied) do not accurately differentiate between items that were studied with normatively effective versus normatively ineffective strategies (Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Shaughnessy, 1981), even when independent measures show that individuals have gained knowledge of differential strategy effectiveness through task experience (Hertzog et al., 2009). In those studies, available knowledge about strategy effectiveness was apparently not accessed when constructing judgments of learning (JOLs) while studying a second list of word pairs, even though differential strategy effectiveness had been experienced on the preceding recall test.

Other evidence suggests that learners may not access beliefs about the contribution of study-test repetitions to memory when making predictions about future learning (Kornell & Bjork, 2009). When different groups of people study an item once and then are asked to predict the likelihood they will recall that item after 1, 2, 3, or 4 study trials (termed predictions of learning, or POLs), they predict relatively similar levels of recall for items regardless of the number of study opportunities signified for a POL. This and other findings led Kornell and Bjork (2009) to propose that people have a *stability bias* when making POLs (for other evidence see, Koriat, Bjork, Sheffer, & Bar, 2004; Kornell, Rhodes, Castel, & Tauber, 2011).

Understanding why the stability bias occurs and how to overcome it could have important educational implications. A stability bias would be troubling if it reflected inaccurate knowledge about study effectiveness. Students might opt not to study material more than once if they believe that additional study will have a minimal influence on later memory. On the other hand, if the stability bias merely reflects insensitivity of predictions to the number of study opportunities because that variable was not accessed or considered when making a POL, then it might have little practical consequence for student performance.

Do people simply fail to use their beliefs about how study improves learning when predicting their memory? If so, it should be possible to manipulate accessibility of the beliefs and therefore manipulate the sensitivity of POLs to the number of study-test trials. One potentially important factor is the method of framing the judgment. It is well known that within-subjects manipulations that help frame differences in levels of an independent variable are often needed to reveal experimental effects on metacognitive judgments (e.g., Carroll & Nelson, 1993; Koriat et al., 2004). However, Kornell and Bjork (2009) found that the stability bias was observed even with such a within-subject manipulation. Participants were instructed that they would receive 4 blocks of study-test practice for 28 paired associates and then gave POLs. Specifically, they were asked immediately after studying each pair for the first time (i.e., during the first block of study) to predict how likely they were to remember that pair on subsequent tests, either test 1, test 2, test 3, or test 4. Surprisingly, people predicted that they would recall a similar percentage of items on each of the four tests and this stability bias persisted even when participants were instructed not to undervalue the contribution of repeated study on their memory.

One possible explanation of this stability bias is that the framing of the judgment task used by Kornell and Bjork (2009) did not activate beliefs about the contribution of increasing study opportunities to learning. For any given item, participants were asked to predict their performance for either test 1, test 2, test 3, or test 4. Thus, the quantity emphasized by the POL framing was the number of tests. Given that (a) metacognitive judgments are typically not sensitive to the memorial benefits of retrieval practice (Karpicke, Butler, & Roeder, 2009; Karpicke & Roediger, 2008; Kornell & Son, 2009; Roediger & Karpicke, 2006, Tullis, Finley, & Benjamin, 2013) and (b) people incorrectly report that repeated study is much more effective for learning than repeated testing (Kornell & Son, 2009), people may not believe that testing per se is beneficial for learning. In fact, students typically use practice tests to assess their memory rather than as a technique to foster additional learning (Hartwig & Dunlosky, 2012; Kornell & Son, 2009). Framing predictions to emphasize the number of tests therefore might not activate beliefs about additional learning occurring on each study-test trial. In contrast, if POLs were framed to emphasize the number of study opportunities that would occur, it is possible that people would be more likely to access beliefs about how studying improves learning which they could use to make their predictions.

Although, it is currently unclear whether framing POLs to emphasize testing over studying contributes to the stability bias, there is evidence that other types of framing influence the cues learners attend to when making metacognitive judgments in multi-trial learning tasks (Koriat et al., 2004; Serra and England, 2012). For instance, Koriat et al. (2004) manipulated the framing of JOLs to emphasize remembering (e.g., “How likely do you feel you will remember this item?”) or forgetting (e.g., “How likely do you feel you will forget this item?”). Participants were asked to predict performance on a test that would occur immediately, one day, one week, or even 1 year after study. People predicted forgetting would occur across time with forget-framed JOLs. However, remember-framed JOLs were susceptible to a stability bias, demonstrating no or minimal change in predicted memory performance across retention intervals. Forget framing apparently drew people’s attention to the possibility of memory failures occurring across time.

Since Koriat et al.’s (2004) seminal work, others have evaluated how forget framing and remember framing affects metacognitive judgments (Finn, 2008; Halamish, McGillivray, & Castel, 2011; Koriat et al., 2004; Kornell & Bjork, 2009; Rhodes & Castel, 2008; Serra & England, 2012; Tauber & Rhodes, 2012a). Most relevant to the current experiments is work conducted by Kornell and Bjork (2009) that examined whether forget framing would eliminate the stability bias that occurs for POLs for items that would receive one to four study-test opportunities. They found that forget framing failed to eliminate the stability bias – possibly because forget-framed judgments appear to be psychologically drawn to the midpoint of judgment scales (Serra & England, 2012). Nevertheless, it is possible that other types of framing could reduce the stability bias, depending on the degree to which the frame activates beliefs about performance increases during multi-trial learning.

Overview of the Current Experiments

The current experiments evaluated whether the stability bias detected by Kornell and Bjork (2009) could be attributed to deficient belief content – a lack of awareness of the benefits of multiple study opportunities – or to a failure to access accurate beliefs about learning in multiple study contexts when making POLs. Kornell and Bjork (2009) defined the stability bias as, “A failure to predict the degree to which memory can change over time (pg. 450).” To remain consistent, we operationally define the stability bias in the same manner by assuming it reflects an underestimation of learning rates relative to the rate of actual memory changes across time.

In Experiment 1, we tested the accessibility hypothesis for the stability bias by manipulating the framing of POLs to emphasize either the number of tests or the number of study opportunities that would occur. We predicted that framing POLs to emphasize the number of future study opportunities that would occur for an item would activate beliefs about how study improves memory that they would incorporate into their predictions, reducing the stability bias relative to POLs given a test-based framing. In Experiment 2, we evaluated an alternative explanation for our hypothesized framing effects that arose from our wording of the frames in Experiment 1. Namely, we evaluated whether framing effects stem from highlighting quantity in the frame opposed to activating study beliefs.

In Experiments 3 to 5, we empirically separated memory belief content from belief accessibility when making POLs by explicitly and independently measuring pre-experimental beliefs about learning. To measure beliefs about the contribution of study repetitions to memory, we asked participants before studying items to make global predictions about the number of pairs they believed they could recall after 1, 2, 3, and 4 study opportunities. Participants made these predictions in Experiments 3 and 4 using a visual analog rating scale that allowed them to draw a subjective learning curve to indicate their predicted rate of learning across study-test trials. In Experiment 5, they made similar predictions using numerical scales in which they reported the number of pairs they expected to recall on each trial. We used these responses to characterize persons with respect to the degree of general expected performance improvement during multiple study-test trials. We hypothesized that our participants would generate robust predicted learning curves, manifested in nonzero slopes of expected change across study-test trials. A critical advantage of this method is that we could use their responses to estimate within-person slopes of predictions over trials, and then use these fitted slopes to examine the relationship between beliefs in learning and analogous slopes generated from their averaged item-level POLs. If the stability bias is due to a categorical failure to access beliefs about study repetition benefits, then individual differences in the predicted rate of learning from global learning curves should be independent of (i.e., not correlate with) item-level POLs. However, if POLs merely discount the effects of learning yet are sensitive to belief-based processes, the rate of predicted learning measured by pre-study global judgments should be correlated with the rate of predicted learning for item-level POLs.

Experiment 1

Method

Participants—Fifty-two undergraduates from the Georgia Institute of Technology participated in this experiment for course credit in introductory psychology. Participants were randomly assigned to either the test frame ($N = 26$) or study frame group ($N = 26$).

Materials—The same paired associates that were classified as being difficult to learn by Kornell and Bjork (2009) were used in the current experiment. These materials consisted of 60 unrelated paired associates (e.g., CLEMENCY – IDIOM).

Procedure—The procedure was modeled after Kornell and Bjork (2009, Experiment 6). Participants were instructed that they would be learning paired associates (e.g., DOG-SPOON) and that they would be tested for their memory of the second word of each pair after being cued by the first word of that pair. Participants first completed a practice phase to familiarize them with the material and procedure of the experiment. Ten paired associates were presented individually for 4 seconds each. After all the pairs were presented a cued recall test was administered. During this test, the cue word of each pair was presented (e.g., DOG-?) and participants were prompted to type the target word. After this practice phase concluded, participants were instructed that they would now study 60 new paired associates. They were told that they would study each pair 4 times and be tested on each pair 4 times. They were also instructed that during the first block of study-test trials, they would make POLs for each pair. For POLs, they would have to make a prediction of the likelihood they would remember the second word of the pair just studied on test 1, test 2, test 3, or test 4. They were instructed that they would only make one test prediction for each pair and the instructions emphasized that they would have an opportunity to study each pair once before each test. That is, all pairs would be studied once by test 1, twice by the time of test 2, three times by test 3, and four times by test 4.

During the study phase of each trial, each pair was presented individually for 4 seconds. Immediately after the presentation of a pair, participants were prompted to make a POL for test 1, test 2, test 3, or test 4. Participants in the study frame group were asked, “How likely are you to remember the second word of the previous pair after the number of study opportunities highlighted below?” An example of a POL trial for the study frame group is presented in Figure 1. Four boxes were presented corresponding to each test trial labeled 1, 2, 3, and 4. However, 3 of the 4 boxes were grayed out. A blinking cursor was placed in the remaining box to highlight the target test for their prediction (in this case test 3). Participants made predictions for each test an equal number of times (15 each) and the test trial that a POL was made for a particular item was always randomly assigned. The POL procedure for participants in the test frame group was identical to the study frame group with the exception that they were asked, “How likely are you to remember the second word of the previous pair during the test opportunity highlighted below?”

After studying all 60 word pairs and making POLs for each one, participants were given a cued-recall test. Following testing the procedure above repeated for 3 additional study-test trials. However, participants were not asked to make POLs on any of the remaining trials.

Results & Discussion

The mean percentage of word pairs correctly recalled and the mean POL for tests 1 to 4 are presented in Figure 2 as a function of framing group. First consider actual recall performance in Figure 2. Recall increased from test 1 to test 4 for both the test frame (top panel) and study frame groups (bottom panel). Consistent with this observation a 2(framing: test vs. study) $\times$ 4(test: 1, 2, 3, or 4) repeated measures ANOVA with a planned comparison for linear trend (see Hertzog, 1994) revealed a linear effect for test trial, $F(1,50) = 182.85$, $MSE = 36211.02$, $p < .001$, $\eta_p^2 = .79$. Effects for group, $F(1,50) = .17$, $MSE = 225.70$, $p = .38$, $\eta_p^2 = .003$, and the group $\times$ linear trend interaction, $F(1,50) = .05$, $MSE = 8.96$, $p = .83$, $\eta_p^2 = .001$, were not significant. Recall increased across trials for both groups and prediction framing did not influence memory performance.

Next consider predicted recall in Figure 2 (POLs). Participants in the test frame group predicted less learning across trials than participants in the study frame group. A 2(framing) $\times$ 4(trial) repeated measures ANOVA revealed a linear effect for Trial $F(1,50) = 53.77$, $MSE = 13890.32$, $p < .001$, $\eta_p^2 = .52$, and no main effect for framing, $F(1,50) = .47$, $MSE = 556.29$, $p = .50$, $\eta_p^2 = .01$.

Consistent with our hypothesis about study framing versus test framing, the Framing $\times$ Trial interaction was significant, $F(1,50) = 8.96$, $MSE = 2315.17$, $p < .01$, $\eta_p^2 = .15$. The reliable interaction reflected greater POL slopes for the study frame group ($M = 10.29$, $SE = 1.50$) than the test frame group ($M = 4.33$, $SE = 1.32$), $t(50) = 2.99$, $p < .01$, $d = .83$. Thus, people predicted a higher rate of learning with each additional study-test trial when POLs were framed to emphasize study opportunities relative to test experience. On the other hand, the rate of actual learning (recall slope) did not differ for the study ($M = 11.62$, $SE = 1.34$) or test frame groups, ($M = 11.99$, $SE = 1.12$), $t(50) = -0.21$, $p = .82$, $d = .06$, consistent with the previous recall analysis.

To evaluate explicitly the relationship between actual performance and predicted performance, we computed slopes for each and then computed difference scores (POL slope – Recall slope) to estimate the degree of under/over confidence in the rate learning generated by POLs. All slopes and difference scores were computed for each individual and means reflect means across individuals' values. This slope deviation measure revealed that the test frame group significantly under-predicted the amount of actual learning that would occur in the task ($M = -7.66$, $SE = 1.71$), $t(25) = 4.48$, $p < .001$, $d = .88$, whereas the study frame group's deviation measure did not significantly differ from zero ($M = -1.32$, $SE = 1.55$), $t(25) = 0.86$, $p = .40$, $d = .17$. Thus, test frame predictions were characterized by a stability bias in which people under-predicted the amount of actual learning that would occur in the task, but study frame predictions were sensitive to the rate of actual learning that would be generated by multiple study-test exposures.

Experiment 2

In Experiment 1, POLs were sensitive to the number of study-test trials when predictions were framed to emphasize study but not testing. However, there is a potential confound in Experiment 1 because the wording of the frames may have drawn peoples' attention to the

number of trials that would occur in the study frame but not the test frame group. For the study frame, people were asked “How likely are you to remember the second word of the previous pair after *the number* of study opportunities highlighted below?” In contrast, people in the test frame group were asked, “How likely are you to remember the second word of the previous pair during the test opportunity highlighted below?” The study frame group explicitly mentioned number of opportunities in the frame, but the test frame did not. It is possible that the explicit framing of the number of learning opportunities for pairs in the study frame group influenced POLs and not the study frame itself.

The goal of Experiment 2 was to address this confound by manipulating explicit versus implicit framing of the number of opportunities that would occur for study and test framed POLs. If study framing is responsible for the sensitivity of POLs to new learning across trials, then POLs will predict more learning across time when they are framed in terms of study than when they are framed to emphasize testing (as in Experiment 1) regardless of whether number is explicitly mentioned in the frame. However, if mentioning quantity in the frame is responsible for the sensitivity of POLs to new learning across trials, then emphasizing quantity with an explicit number frame will influence POLs in both the testing and study frame groups.

Method

Participants—Ninety undergraduates from Georgia Institute of Technology participated in this experiment for course credit in introductory psychology. Participants were randomly assigned to either the test frame with quantity emphasized group ($N = 23$), the test frame with quantity not emphasized group ($N = 23$), the study frame with quantity emphasized group ($N = 22$), or the study frame with quantity not emphasized group ($N = 22$).

Materials and Procedure

The materials and procedure were identical to Experiment 1 with the exception of the prompts used to elicit POLs. Participants in the test frame with quantity emphasized group were asked, “How likely are you to remember the second word from the previous word pair during the test number highlighted below?” Participants in the test frame with quantity not emphasized group were asked, “How likely are you to remember the second word from the previous word pair during the test opportunity highlighted below?” Participants in the study frame with quantity emphasized group were asked, “How likely are you to remember the second word from the previous word pair after the study number highlighted below?” Finally, participants in the study frame with quantity not emphasized group were asked, “How likely are you to remember the second word from the previous word pair after the study opportunity highlighted below?”

Results & Discussion

The mean percentage of word pairs correctly recalled and the mean POL for tests 1 to 4 are presented in Figure 3 as a function of learning and quantity framing. We collapsed across the mean percentage of words recalled for quantity framing groups in Figure 3 because framing did not influence recall performance. A 2(Learning frame: study vs. test) $\times$ 2(quantity frame: quantity emphasized vs. not emphasized) $\times$ 4 (test: 1, 2, 3, or 4) ANOVA

revealed only a significant linear trend effect for test trial, $F(1,86) = 371.56$, $MSE = 74781.98$, $p < .001$, $\eta_p^2 = .81$. Effects for learning frame, $F(1,86) = 1.05$, $MSE = 1493.06$, $p = .31$, $\eta_p^2 = .01$, quantity frame, $F(1,86) = .01$, $MSE = .37$, $p = .99$, and all interaction effects were not significant, $F_s < 2.15$, $\eta_p^2 < .03$.

Concerning predicted performance in Figure 3 (POLs), participants predicted more learning across time when POLs were framed in terms of study than when POLs were framed in terms of testing. Most important, quantity framing did not influence POLs for either learning frame. Consistent with these observations, a 2(Learning frame: study vs. test) $\times$ 2(quantity frame: quantity emphasized vs. not emphasized) $\times$ 4 (test: 1, 2, 3, or 4) ANOVA revealed only a significant linear trend effect for test trial, $F(1,86) = 23.22$, $MSE = 4277.64$, $p < .001$, $\eta_p^2 = .21$, and a Trial $\times$ Learning frame interaction effect, $F(1,86) = 8.10$, $MSE = 1492.58$, $p < .05$, $\eta_p^2 = .09$. Effects for learning frame, quantity frame, and all other interaction effects were not significant, $F_s < 1$. The significant Trial $\times$ Learning frame interaction occurred because POL slopes were greater for the study frame ($M = 4.91$, $SE = .98$), than the test framed groups ($M = 1.26$, $SE = .82$), $t(88) = 2.87$, $p < .01$, $d = .61$. The study framed POL slopes differed significantly from zero, $t(43) = 5.03$, $p < .001$, $d = .76$, whereas the test framed POL slopes did not, $t(45) = 1.55$, $p = .13$, $d = .23$. In contrast, the rate of actual learning (recall slopes) did not differ for study ($M = 12.64$, $SE = .85$) or test frame POLs ($M = 13.15$, $SE = 1.02$), $t(88) = .38$, $p = .71$, $d = .08$.

We again computed slope deviation scores to evaluate under/over confidence in the rate of predicted learning generated by POLs. Consistent with findings from Experiment 1, people underestimated the amount of actual learning that would occur when POLs were framed in terms of testing ($M = -11.89$, $SE = 1.08$) than when they were framed in terms of study ($M = -7.73$, $SE = 1.12$), $t(88) = 2.66$, $p < .01$, $d = .56$. However, unlike Experiment 1, POLs under-predicted actual learning even when they were framed to emphasize study. This finding was characterized by significant non-zero deviation scores for both study, $t(43) = 6.88$, $p < .001$, $d = 1.62$, and test framed POLs, $t(45) = 10.98$, $p < .001$, $d = 1.04$.

In summary, people predicted more learning across trials when POLs were framed in terms of study than when they were framed in terms of testing. However, quantity framing did not influence the rate of predicted learning across trials. Taken together, these results suggest that the effects for study framing in Experiment 1 were not due to emphasizing quantity in the frame. Instead the effects of study framing on POLs likely originate from study framing activating beliefs about learning across time that test framing fails to activate.

Experiment 3

Experiment 1 and 2 revealed that POLs may be sensitive to beliefs about the contribution of repeated study to learning when judgments are framed to activate these beliefs. The main goal of Experiment 3 was to directly measure beliefs about multi-trial study effects on learning independently from (and before) assessing item-level POLs, enabling an empirical evaluation of the relationship of prior beliefs about learning on POLs. Kornell et al. (2011; Experiment 3) showed that persons given descriptions of a two-trial learning study expected better performance on the second study-test trial. However, they did not explicitly link this

belief in learning to the stability bias in POLs found in their other experiments by measuring both variables in the same persons in the same experiment.

We measured beliefs about study benefits using new scaling technique inspired by the General Beliefs about Memory Instrument (Lineweaver & Hertzog, 1998). That questionnaire measures implicit beliefs about the effects of aging on memory by using visual analog rating scales that allow people to draw hypothetical memory change trajectories over the adult life span. Visual analog scales often outperform discrete numerical scales in measurement contexts where detection of small changes is important (Guilford, 1954; Bellamy, Campbell, & Syrotuik, 1999; Brunier & Graydon, 1996; Reips & Funke, 2008). Thus, visual analog scales may be superior for measuring expected learning curves gained from multiple study-test trials on a single list of items. We also used a variation of this graphic rating scale method for item-level POLs to keep measurement scales consistent across the two tasks.

We also measured beliefs about multi-trial study benefits after the learning experiment in order to assess updated knowledge about learning based upon actual task experience. First, after the task description and instructions, but prior to studying any items, individuals predicted levels of performance at each of the four study-test trials using visual analog scaling (see Figure 4). They repeated this set of ratings at the end of the experiment. These ratings captured pre-experimental and post-experimental beliefs about aggregate (list-wide) increases in memory performance across trials.

Experiment 3 also manipulated the number of predictions that people made for each pair. In Experiments 1 and 2, participants made a POL for only one test (1, 2, 3, or 4) for each paired-associate item, with the test that was assessed chosen randomly for each item. We used this method again in Experiment 3, as one level (separate prediction group) of a between-subjects manipulation. However, the experiment also included a group that required POLs to be made for every test for every item (joint prediction group). That is, after studying a word pair for the first time, this group drew a learning curve for that pair to portray their predicted rate of learning for it. We hypothesized that the joint prediction group would be even more likely to consider the influence of repeated study on memory because they were required to jointly evaluate all levels of future study when making POLs for each item.

Method

Participants—Sixty undergraduates from Georgia Institute of Technology participated in this experiment for course credit in introductory psychology. Participants were randomly assigned to either the separate prediction ($N = 31$) or joint prediction group ($N = 29$).

Materials and Procedure—The materials and procedure were similar to Experiment 1. After receiving instructions about the experiment (but prior to beginning it), participants were asked to make a global prediction about the number of pairs they believed they could remember after 1, 2, 3, and 4 study opportunities using the visual analog rating scale (Figure 4). Participants drew a learning curve by clicking a point between 0 and 60 (the number of items on the list) on each of the lines labeled 1, 2, 3, and 4 to indicate their predicted

memory on each study-test block. As is standard with visual analog scales, participants were not given any numerical information about the location of their selections. However, when they clicked a point on the scale, a black circle was presented to mark the location of their selection. They could alter that point by either clicking on a different part of the rating scale or by dragging the circle up or down the vertical rating scale. When they were satisfied with the scale location, they clicked a different button to proceed.

After making the global prediction, participants completed the same task outlined in Experiment 1 with the exception that their POLs were made using the visual analog rating scale. For POL trials, participants were asked “How likely are you to remember the second word of the previous pair after the number of study opportunities highlighted below”. Four vertical lines were presented labeled 1, 2, 3, and 4 to indicate the number of elapsed study opportunities for the target prediction. The top of each line was labeled very likely and the bottom was labeled very unlikely. For the separate prediction group, 3 of the 4 lines were grayed out to indicate the target trial for the POL (as in Experiment 1 and 2). The target trial was always randomly selected with the constraint that an equal number of predictions occurred for each trial. Participants made their prediction by using a mouse to click on any point on the line between the labeled scale endpoints of “very likely” and “very unlikely.” When participants clicked on a point, a small black circle appeared to mark the location of their selection. The method for scaling POLs were similar for the joint prediction group with the exception that all four lines were visible (none were grayed out) and participants made four predictions for each pair by clicking on all four lines.

After completing the experiment, participants made post-study global judgments using the same visual analog rating scale used for pre-study global judgments (Figure 4). However, for post-study global judgments, they were instructed to imagine that they were going to complete the same experiment again, but with new paired associate items.

Results & Discussion

Recall, POLs, and Global Predictions—For POLs and global predictions, the spatial locations of participants’ selections on each visual analog scale were transformed to numerical values between 0 and 100 representing the percentage distance traversed from the bottom to the top of the scale. The bottom end of each scale was assigned a value of 0 and each 4.5 pixel increase from it corresponded to a 1% increase in predictions. We chose to analyze global predictions as percentages (opposed to the number of pairs predicted out of 60) so that they could be directly compared to POLs and participants’ percent recall.

The mean percentage of pairs actually recalled, POLs, pre-study global prediction, and post-study global prediction are presented in Figure 5 for each test for the separate (top panel) and joint prediction (bottom panel) groups. First consider actual recall performance in Figure 5 (solid black lines). A reliable Trial linear trend revealed that recall increased across tests trials for both groups, $F(1,58) = 330.65$, $MSE = 55285.71$, $p < .001$, $\eta_p^2 = .85$. As expected given random assignment, recall did not differ between groups, $F < 1$, $\eta_p^2 = .003$, and there was no hint of a Group X linear Trial interaction, $F < 1$, $\eta_p^2 = .002$.

Next consider POLs presented in Figure 5 (dotted black lines). The linear effect for trial was significant, $F(1,58) = 121.12$, $MSE = 9946.22$, $p < .001$, $\eta_p^2 = .68$, with people predicting improvements in recall across trials. Unexpectedly, we saw no effect of POL format, with no reliable effects for prediction group, $F(1,58) = 1.77$, $MSE = 1976.83$, $p = .19$, $\eta_p^2 = .03$, nor a Group X linear Trial interaction, $F(1,58) = 0.28$, $MSE = 22.62$, $p = .60$, $\eta_p^2 = .01$. Predictions were sensitive to the influence of study repetitions on memory regardless of whether separate predictions were made for each trial or multi-trial predictions were made for each item.

Lastly consider pre-study and post-study global predictions, which reflect pre-experimental beliefs about the influence of multiple study opportunities on memory and changes in those beliefs due to task experience. Linear Trial effects were significant for both pre-study, $F(1,58) = 187.75$, $MSE = 39715.50$, $p < .001$, $\eta_p^2 = .76$, and post-study global predictions, $F(1,58) = 104.37$, $MSE = 34192.01$, $p < .001$, $\eta_p^2 = .64$. Effects for group were not significant for either pre-study global, $F(1,58) = 0.13$, $MSE = 130.93$, $p = .72$, $\eta_p^2 = .002$, or post-study global predictions, $F(1,58) = 2.45$, $MSE = 2953.25$, $p = .12$, $\eta_p^2 = .04$. The associated interactions were also not significant for pre-study, $F(1,58) = 0.39$, $MSE = 82.42$, $p = .54$, $\eta_p^2 = .01$, or post-study global predictions, $F(1,58) = 0.12$, $MSE = 38.94$, $p = .73$, $\eta_p^2 = .002$. If one were to attend merely to the means, it would appear that there was essentially no change in beliefs about learning in multi-trial study-test contexts, perhaps because there was a relatively good correspondence between pre-experimental beliefs about learning and the magnitude of actual performance improvements in the task.

In sum, people believed that memory would improve across study-test trials before and after completing the experiment regardless of prediction group¹. Although this was also true of POLs, they manifested lower slopes than memory predictions and actual memory recall, consistent with a stability bias.

Individual Differences in slopes for POLs, Global Predictions, and Recall—

Given that effects for prediction group were not significant for any variable, we collapsed across groups when examining individual differences in the rates of predicted learning and actual recall. Participants' mean fitted slopes for their POLs, actual recall performance, pre-study global predictions, and post-study global predictions slopes are presented in Table 1. The mean POL slope, $t(59) = 11.06$, $p < .001$, $d = 1.43$, recall slope, $t(59) = 18.32$, $p < .001$, $d = 2.36$, prestudy global prediction slope, $t(59) = 13.76$, $p < .001$, $d = 1.76$, and post-study global prediction slope, $t(59) = 10.29$, $p < .001$, $d = 1.33$, were all significantly different from zero. Note, however, that similar to Experiment 2, participants' POLs under-predicted the rate of actual learning that would occur, $t(59) = 10.73$, $p < .001$, $d = 1.38$.

We used the relationship between slopes to examine the influence of pre-experimental beliefs on POLs. Correlations between slopes are presented in Table 2. Despite the apparent

1. An alternative explanation for the similar POLs for the joint and separate prediction groups is that the prestudy global predictions (which were joint predictions) may have activated beliefs about study benefits for both groups. We evaluated this hypothesis in a separate experiment that manipulated prediction type (separate vs. joint) for POLs and whether participants made prestudy global predictions. The results were consistent with the current findings. There were no differences between POLs for the joint and separate prediction groups regardless of whether people made prestudy global predictions. These data are available from the first author upon request.

stability bias in the mean POLs, POL slopes were most strongly correlated with pre-study global prediction slopes. This finding is consistent with the hypothesis that people's beliefs about study repetition benefits on memory influences expectations about the amount of learning that will occur across study-test trials. Despite the similarity in the slopes for global predictions before and after task experience, post-study global prediction slopes were most strongly related to recall slopes. This finding suggest that individuals alter their expectations about future learning based on monitoring their previous recall performance, with some persons lowering, but others raising their estimated slopes.

Regression analyses were conducted to further evaluate the relationships between beliefs about learning and the rate of predicted and actual learning that occurred in the task. First, we examined whether participants predicted rate of learning for items is influenced by their beliefs about multi-trial study benefits on memory, regressing POL slopes on the pre-study global prediction slopes. The model predicted 40% of the variability in the participants' POL slopes, $R^2 = .40$, $adjusted\ R^2 = .39$, $F(2, 58) = 38.53$, $p < .001$. The pre-study global prediction slopes were positively associated with POL slopes, ($\beta = .63$, $p < .001$).

Next, we examined whether POL slopes and pre-study global predictions slopes predicted the actual rate of learning in the task (recall slopes). The model predicted 14% of the variance in the rate of recall increases across trials, $R^2 = .14$, $adjusted\ R^2 = .11$, $F(3, 57) = 4.62$, $p < .05$, and this effect was predominantly due to effects of POL slopes ($\beta = .38$, $p < .05$). Pre-study global prediction slopes ($\beta = -.01$, $p = .95$) did not predict recall slopes independently of POL slopes. Apparently, POLs, which were grounded in the study of specific items, statistically mediated the relationship of pre-study beliefs on recall performance. This relationship occurred despite the fact that prediction slope means were more similar to actual recall slope means than POL slope means, reinforcing the point that individual differences results need not correspond to patterns seen in average experimental effects across conditions.

Finally, we examined whether people updated their knowledge about the influence of study repetitions on memory after completing the experiment by computing a model regressing post-study global prediction slopes on POL type, pre-study global prediction slopes, POL slopes, and recall slopes. This model predicted 46% of the variability in post-study global prediction slopes, $R^2 = .46$, $adjusted\ R^2 = .43$, $F(4, 56) = 15.61$, $p < .001$. Most important, participants' recall slopes independently predicted post-study global predictions slopes ($\beta = .54$, $p < .001$), controlling on pre-study beliefs and POLs. This outcome suggested that participants monitored their performance and altered their expectations of future learning based on this monitoring. Participants' pre-experimental beliefs about study repetition benefits were also related to their post-study beliefs, ($\beta = .38$, $p < .01$). Thus, there was some stability in individual differences in beliefs about learning as manifested by global performance predictions. POL slopes ($\beta = -.10$, $p = .48$) did not yield unique prediction of post-study beliefs.

As an aside, Kornell and Bjork (2009) routinely contrasted stability in their mean POLs across test trials with subsequent higher levels of recall and performance postdictions. The latter variable probably captures influences of monitoring test performance (e.g., Hertzog,

Price, & Dunlosky, 2008) on perceived learning after the fact. Given the emergent sensitivity to actual recall improvements in the post-study beliefs measure in this experiment, we argue that patterns of postdictions are not ideal for either scaling the stability bias in POLs or capturing beliefs about learning that influence the POLs that operate prior to recall test experience.

Experiment 3 conclusively demonstrated that pre-experimental beliefs about learning influence POLs, that POL slopes subsequently predict recall benefits from multiple study-test opportunities, and that beliefs about learning can be altered by actual task performance. Taken together, these results blunt any concern that POLs are insensitive to pre-experimental beliefs about learning due to a stability bias. Although mean POL slopes assessed with the visual analog rating scale method showed some degree of underestimation in actual recall slopes (that can be construed as a mild stability bias), the individual differences in slopes revealed clear connections between pre-experimental beliefs about multiple-trial learning and POLs.

Considering (a) that the POL slopes were not influenced by explicitly requiring judgments across the 4 study-test trials, relative to the condition where only a single POL was collected, and (b) the relationship between pre-experimental prediction slopes and POL slopes, we infer that beliefs about learning are accessed when making POLs, but that these beliefs are discounted relative to other available cues when scaling confidence, resulting in an underestimation of the learning effect.

Experiment 4

Experiment 3 directly linked peoples' beliefs about learning across multiple study-test trials to their predicted rate of learning for items using a new scaling technique. However, prior beliefs about learning were only examined for people who made study-framed POLs. The goal of Experiment 4 was to replicate the test versus study framing effect demonstrated in Experiment 1 and 2 and to evaluate relations of beliefs about learning across study-test blocks to POLs (as in Experiment 3) for participants who made either study-framed or test-framed POLs. If we assume that beliefs about learning are accessed when making POLs, regardless of framing, then one could expect reliable correlations of prediction slopes with POL slopes in both conditions, despite differences in mean POL slopes. However, if test framing restricts or eliminates access to beliefs about learning, one would expect test framing to significantly reduce or even eliminate the correlation of global prediction slopes with POL slopes for test-framed POLs. In other words, if test-framed POLs do not activate beliefs about study benefits, then peoples' beliefs about learning should diverge from their item-level POLs in the test framed group. Although people randomly assigned to study or test framed groups would be equally likely to expect learning to improve across study-test blocks, only study-framed POLs would demonstrate sensitivity to expected performance improvements across trials.

Method

Participants—Fifty-six undergraduates from the Georgia Institute of Technology participated in this experiment for course credit in introductory psychology. Participants were randomly assigned to either the test frame ($N = 27$) or study frame group ($N = 29$).

Materials and Procedure—The materials and procedure were identical to the separate prediction groups from Experiment 3 with the exception that a second group was also included in which POLs were test framed. Test frame POLs were scaled in the same manner as study framed POLs using the visual analog rating scale described in Experiment 3. The prompts for the study and test frame POLs were identical to the quantity emphasized study and test frame groups from Experiment 2. All prestudy global predictions were study framed (see Figure 4).

Results & Discussion

Recall, POLs and Global Predictions—The means for participants' percentage recall, POLs, prestudy global predictions, and post-study global predictions are presented in Figure 6 for the test frame (top panel) and study frame (bottom panel) groups. Separate 2(framing) $\times$ 4(test trial) ANOVAs were computed for each variable.

First consider actual recall performance in Figure 6. Consistent with findings from previous experiments, recall improved across trials, $F(1,54) = 203.23$, $MSE = 43120.20$, $p < .001$, $\eta_p^2 = .79$, and framing did not influence actual recall performance, $F(1,54) = 1.15$, $MSE = 1573.63$, $p = .29$, $\eta_p^2 = .02$. The trial $\times$ framing interaction was also not significant, $F(1,54) = 1.25$, $MSE = 264.84$, $p = .27$, $\eta_p^2 = .02$.

The critical data concern POLs in the two framing conditions. As before, framing influenced the mean rate of predicted learning in POLs. The linear trend effect for trial, $F(1,54) = 13.44$, $MSE = 1471.96$, $p < .001$, $\eta_p^2 = .20$, and an effect for Framing, $F(1,54) = 5.41$, $MSE = 4322.07$, $p < .05$, $\eta_p^2 = .09$, were both significant. Most important, these effects were qualified by a significant Trial $\times$ Framing Interaction Effect, $F(1,54) = 5.10$, $MSE = 557.99$, $p < .05$, $\eta_p^2 = .09$. This interaction occurred because people again predicted a higher rate of learning across trials when POLs were study framed than when they were test framed (see POL slopes in Table 1), $t(54) = 2.26$, $p < .05$, $d = .61$.

Concerning global predictions, only the significant linear trend effect for test trial was significant for prestudy $F(1,54) = 72.59$, $MSE = 23118.15$, $p < .001$, $\eta_p^2 = .57$, and post-study global predictions $F(1,54) = 117.20$, $MSE = 28983.93$, $p < .001$, $\eta_p^2 = .69$. Effects for framing and the trial $\times$ framing interaction effect were not significant for either variable, $F_s < 1.98$. As can be seen in Figure 6, experience in the task lowered prediction intercepts but did not materially affect the global prediction slopes. The lack of a pre-experimental difference in beliefs merely reflects successful random assignment; the lack of a post-task difference suggests that framing did not alter the influence of monitoring task learning on post-experimental beliefs about how much learning would occur across each study-test trial. Thus, people in the test frame group expected learning to occur across trials, but their POLs did not accurately reflect this belief.

The means of participants' fitted slopes for their POLs, actual recall performance, pre-study global predictions, and post-study global predictions are presented in Table 1. Only POL slopes differed as a function of framing, $t(54) = 2.26, p < .05, d = .61$ (all other slopes, $ts < 1.5$). For the study frame group, prestudy global prediction slopes, $t(28) = 9.47, p < .001, d = 1.76$, POL slopes, $t(28) = 3.36, p < .01, d = .68$, recall slopes, $t(28) = 10.43, p < .001, d = 1.93$, and post-study global prediction slopes, $t(28) = 6.99, p < .001, d = 1.30$, all differed significantly from zero. For the test framed group, POL slopes did not significantly differ from zero, $t(26) = 1.27, p = .22, d = .24$, but prestudy global prediction slopes, $t(26) = 4.09, p < .001, d = .79$, recall slopes, $t(26) = 9.83, p < .001, d = 1.89$, and post-study global prediction slopes, $t(26) = 9.34, p < .001, d = 1.80$, did significantly differ from zero. Thus, the pattern of POLs replicated Experiment 2, despite the use of a visual analog scale instead of numeric estimates. For both types of framing, POLs underestimated actual learning rates, but only study-framed POL slopes were reliably greater than 0.

Individual Differences in slopes for POLs, Global Predictions, and Recall—

Correlations and regression analyses were conducted on the fitted slopes presented in Table 1 to examine the influence of beliefs on POLs. These correlations are presented in Table 2 as a function of framing group. As in Experiment 3, pre-experimental global prediction slopes were reliably correlated with POL slopes for study framed POLs. The analogous correlation for test-framed POLs missed significance ($p = .07$), but the correlation was in the expected positive direction. Caution is warranted in interpreting this relationship given our limited sample size. A Fisher's r -to- z test revealed that these correlations were not reliably different from each other, $z = .29, p = .77$.

A model regressing POL slopes on framing group and prestudy global predictions accounted for variance in the participants' POL slopes, $R^2 = .20, adjusted R^2 = .17, F(2, 53) = 6.47, p < .01$. Pre-study global prediction slopes were also positively associated with POL slopes, ($\beta = .34, p < .01$), and framing group was also associated with POL slopes, but this later relationship was only marginally significant, ($\beta = -.23, p = .07$). Adding a framing type X prestudy global prediction slopes interaction term ($\beta = -.72, p = .13$) to this model ($R^2 = .23, adjusted R^2 = .19, F(3, 52) = 5.19, p < .01$), did not significantly increase the model fit, $R^2 = .03, F(1, 52) = 2.31, p < .14$. In summary, participants' beliefs about multi-trial learning exerted a strong influence on their rate of predicted learning for items, and it appeared that this relationship was present for both framing groups.

Next, we examined whether framing group, POL slopes, and pre-study global predictions slopes predicted the actual rate of learning in the task (recall slopes). The model was not significant, $R^2 = .10, adjusted R^2 = .04, F(3, 52) = 1.85, p = .15$. POL slopes ($\beta = .15, p = .31$), pre-study global prediction slopes ($\beta = .19, p = .19$) and framing group ($\beta = -.07, p = .61$) did not predict recall slopes.

Finally, we examined changes in performance expectations by computing a model regressing post-study global prediction slopes on framing group, pre-study global prediction slopes, POL slopes, and recall slopes. This model predicted 53% of the variability in post-study global prediction slopes, $R^2 = .53, adjusted R^2 = .50, F(4, 51) = 14.47, p < .001$. Participants' recall slopes again independently predicted post-study global predictions slopes

($\beta = .68, p < .001$). Participants' pre-study global prediction slopes, ($\beta = .10, p = .34$), POL slopes ($\beta = .07, p = .50$), and framing group ($\beta = .05, p = .59$) did not predict post-study global prediction slopes controlling on recall slopes. Adding interaction coefficients to the model ($R^2 = .54$, *adjusted* $R^2 = .48$, $F(7, 48) = 8.11, p < .001$) did not improve its fit, $R^2 = .01$, $F(3, 48) = .36, p < .78$. The framing X prestudy global prediction slopes ($\beta = -.43, p = .33$), framing X POL slopes ($\beta = .07, p = .84$), and framing X recall slopes ($\beta = -.02, p = .95$) interactions were all not significant. These results indicate that participants in both groups monitored their task performance and updated their expectations about future learning based on this monitoring.

In summary, the outcomes from Experiment 4 support the argument that beliefs about learning are accessed when making POLs. Test framing does not completely eliminate access to beliefs about learning across study-test trials. Instead Test framing POLs apparently causes participants to discount these beliefs when scaling their POLs across trials, contributing to the apparent stability bias. However, since beliefs about learning were always study framed (see Figure 4), it is possible that belief access during POLs occurred in the test frame group because the prestudy global prediction reminded learners that study would occur on each trial. If so, the current results still suggest that the test framing caused learners to discount these beliefs which resulted in shallower predicted learning slopes than the study framed group.

Experiment 5

Although the experiments reported to this point indicated that framing predictions to emphasize testing accounts for much of the stability bias, it is also clear that study-framing still resulted in an underestimation of actual learning (Experiments 2 and 4). These latter findings are consistent with evidence that a stability bias persists even when predictions are framed to emphasize study repetitions. Kornell et al. (2011) examined the joint influence of font size of words and study repetitions on predictions of learning. Participants viewed words presented in either a large (48 point) or small font (18 point) and predicted the likelihood they would recall that word after 1 or 4 study repetitions. People predict that words presented in a large font are more memorable than words presented in a small font (Rhodes & Castel, 2008). Consistent with this finding, Kornell et al. (2011) found that font size influenced predictions more than the number of study repetitions that would occur, which suggests that the stability bias may be in part due to people overvaluing the effects of stimulus features over effects of study when forecasting memory performance (e.g., Koriat, 1997).

Given our argument that learning beliefs are accessed when making study-framed POLs, why would such effects occur? A cue overshadowing account (Price & Yates, 1993) would argue that the relative salience of perceptual features of the stimuli influences the degree of learning cue discounting. That is, a highly salient cue may draw attention away from other cues and decrease the impact of these cues on estimates of performance. Kornell et al. (2011) appeared to generate an experimental context that would draw attention more to font size than study repetitions. Their font size for large-font words was larger than the font size for text indicating the number of study repetitions that would occur (see Figure 1 from

Kornell et al., 2011). There was also less environmental support in the POL prompt relative to previous experiments to remind people about the presence of multiple study-test trials. In the current experiments and Kornell and Bjork (2009), the location where people provided their predictions on the computer screen was spatially associated with the study-test trial that was the target of their prediction (see Figure 1). However in Kornell et al. (2011), participants made all predictions in the same location on the computer screen. Without the spatial frame to remind people of the presence of multiple study-test trials, people may have been less likely to access the number of study repetitions that would occur when making the POL. If so, they may have overvalued font size when making their predictions because it was a perceptually more salient cue that captured their attention.

A final possible explanation for the differences in experiments is that Kornell et al. (2011) had people make predictions for words studied once on some trials and words that would be studied 4 times on other trials. In the current experiments, people made predictions for pairs that would receive 1, 2, 3, or 4 study opportunities. The contrast between 4 different potential study opportunities (1, 2, 3, and 4) versus 2 different study opportunities (1 versus 4) may have increased the likelihood that learners activated beliefs about multi-trial study benefits on their memories when making predictions.

Experiment 5 evaluated all three of the hypotheses listed above. First, to examine the font-salience hypothesis, we manipulated font size of words during study. Half the word pairs were presented in a large font (48 point) and half were presented in a small font (12 point). If people overvalue stimulus characteristics when making predictions, we hypothesized that font size would influence predictions more than the number of study repetitions, as found by Kornell et al. (2011). Next, to evaluate the cue-overshadowing hypothesis, we manipulated the spatial context of the POLs. One group received a multi-trial spatial context that could potentially remind them about the presence of multiple study-test trials when making POLs. For this group, the spatial location where POLs were made for pairs studied between 1 and 4 times was organized from left to right on the computer screen in the same manner as in Experiments 1 to 4. A second group received no spatial context support. For this group, the spatial location where POLs were made on the computer screen was consistent across trials for pairs that would receive one or more study opportunities (as in Kornell et al., 2011). We hypothesized that spatial cues would remind people about the presence of multiple study test trials, and hence increase the likelihood that they access their beliefs about how study repetitions influence memory when making POLs. If cue overshadowing can explain why people overvalued font size in Kornell et al. (2011), then font size should influence POLs more than study repetitions for the no spatial support group but not for the spatial support group.

Finally, to test the study repetition-contrast hypothesis, we manipulated whether people made predictions for pairs that would be studied once versus four times or pairs that would be studied 1, 2, 3, or 4 times. All groups made only one prediction per trial and the number of study repetitions that would occur for the predicted item was always randomly assigned. If the study repetition contrast during predictions influences belief access, then people who were asked to make POLs for 4 different categories (i.e., pairs studied 1, 2, 3, or 4 times)

should predict more learning across study-test trials than persons making POLs for either 1 or 4 study times.

Given that Kornell et al. (2011) scaled POLs by having persons enter numbers in boxes and we were interested in replicating their findings, we also used numerical scales to elicit POLs in Experiment 5. We again measured beliefs about study repetition benefits on memory before and after the experiment. However, in contrast to Experiments 3 and 4, we used numerical scales to elicit these pre-study beliefs in learning identical to those used for POLs so to keep scaling methods consistent within the task. Doing so also allowed us to examine whether findings in Experiments 3 and 4 regarding beliefs about learning and POLs were specific to the use of a visual analog scale.

Method

Participants—Seventy-eight undergraduates from Georgia Institute of Technology participated in this experiment for course credit in introductory psychology. Participants were randomly assigned to either the no spatial support group who made predictions for pairs studied once or four times ($N = 21$), the no spatial support group who made predictions for pairs studied 1, 2, 3, or 4 times ($N = 17$), the spatial support group who made predictions for pairs studied once or four times ($N = 20$), or the spatial support group who made predictions for pairs studied 1, 2, 3, or 4 times ($N = 20$).

Materials and Procedure—The same materials used in Experiments 1 to 4 were also used in Experiment 5. However, we decreased the number of practice trials available from 10 to 6 so that 4 additional word pairs could be used in the study task. This allowed us to increase the number of observations during the study task to 64 paired associates (opposed to 60). This small increase in stimuli also allowed for an even distribution of stimuli to different prediction types, as specified in the following paragraph. The procedure was similar to Experiment 3 and used only studied framed POLs. Following instructions and practice, participants made pre-study global predictions in the same manner as Experiments 3 and 4 with the exceptions that the visual analog scales depicted in Figure 4 were replaced with text boxes and participants typed values between 0 and 64 to indicate the number of pairs they expected to recall on each study-test trial.

After making pre-study global predictions, participants began the study task. During study, half the word pairs were randomly assigned to be presented in 48 point Arial font and half in 12 point Arial font. The study task was identical to the previous experiments with the exception of how POLs were elicited during the first study-test block. During POLs, people in the spatial support group who made predictions for pairs studied 1, 2, 3, or 4 times viewed a screen similar to Figure 1. The only difference was that the cue word of the previously studied pair was also presented in either a large or small font in the center of the computer screen above the text boxes. A blinking cursor indicated the target test for their predictions. Participants made predictions for each test an equal number of times (16 each, 8 for each font size) and the test trial that a POL was made for a particular item was always randomly assigned.

The POL trials were similar for the spatial group who made predictions for pairs studied once or four times. However, the text boxes for pairs studied 2 and 3 times were removed completely. Again one text box was grayed out (either 1 or 4) and a blinking cursor was presented in the other box. Participants made predictions for pairs that were studied once on half the trials (16 for each font size) and for pairs that would be studied four times (16 of each font size) for the other half. The POL procedure was nearly identical for each corresponding no spatial support group with the exception that only one text box was presented at the center of computer screen (all other text boxes were removed completely). On each trial, this text box was labeled 1, 2, 3, or 4 to indicate the number of study opportunities that would occur for that pair.

After completing all four study-test blocks, participants were asked to make a post-study global prediction about the number of word pairs they expected to recall if they completed the same experiment again with a new set of pairs. Participants typed their responses in the same manner as they did for pre-study global predictions.

Results & Discussion

The mean percentages of word pairs correctly recalled and mean POL for pairs studied once and four times are presented in the left panel of Figure 7 as a function of font size. We collapsed across spatial support and study repetition contrast groups in Figure 7 because neither influenced predictions (no main effects or interactions), all F s < 1. Thus, the current results did not support the cue overshadowing hypothesis or the differential study-repetition hypothesis. For ease of interpretation, we present only analyses of the effects of study repetition and font size below.

First consider actual recall in Figure 7. Mean recall increased across study-test trials, $F(1,77) = 224.94$, $MSE = 7.23$, $p < .001$, $\eta_p^2 = .75$, and people recalled slightly more large font than small font size words, $F(1,77) = 9.16$, $MSE = .12$, $p < .01$, $\eta_p^2 = .11$. The trial X font size interaction was not significant, $F(1,77) = .45$, $MSE = .002$, $p = .51$, $\eta_p^2 = .01$. The effect for font size on recall was surprising because font size typically does not influence memory performance despite consistently influencing JOLs (Rhodes & Castel, 2008; Mueller, Dunlosky, Tauber, & Rhodes, 2014). Interestingly, the only other experiments to examine the influence of font size on POLs (i.e., Kornell et al., 2009), also found enhanced memory for words presented in a large font relative to words presented in a small font. Taken together, these results suggest that POLs may elicit reactive effects on memory that influence how people process words in different fonts in ways that JOLs do not.

For POLs, we analyzed only predictions for pairs studied once and four times because the study repetition contrast groups did not make POLs for pairs studied two or three times. A $2(\text{study repetition: study once vs. four times}) \times 2(\text{font size: large vs. small})$ repeated measures ANOVA revealed an effect for study repetitions, $F(1,77) = 55.55$, $MSE = 12000.72$, $p < .001$, $\eta_p^2 = .42$, and an effect for font size, $F(1,77) = 6.341$, $MSE = 167.71$, $p < .05$, $\eta_p^2 = .08$. The interaction was not significant, $F(1,77) = .002$, $MSE = .06$, $p = .96$, $\eta_p^2 = .00$. In sum, participants' POLs were sensitive to both study repetitions and font size. However as is evident in Figure 7, study repetitions influenced POLs more than font size. Kornell et al. (2011) suggested that the stability bias may be in part due to people

overvaluing ease-of-processing, as captured by font size, when making predictions. The present results did not support this conclusion, instead suggesting that our participants weighed the amount of study an item would receive more than the font size of pairs when making their predictions. One explanation for this difference is that it represents random sampling of effect sizes for the two different independent variables (see Abelson, 1995). Alternatively, people may indeed expect that repeated study influences memory more than font size. If so, it is unclear why Kornell et al. (2011) did not obtain this outcome.

The mean pre-study global prediction and post-study global predictions for each study-test trial are presented in the right panel of Figure 7. The linear trend effect was significant for both pre-study, $F(1,77) = 143.02$, $MSE = 27896.64$, $p < .001$, $\eta_p^2 = .65$, and post-study global predictions, $F(1,77) = 117.39$, $MSE = 16236.67$, $p < .001$, $\eta_p^2 = .60$. These results demonstrate pre-experimental beliefs in benefits of multiple study-test opportunities with a numerical scaling procedure, showing this effect not to be dependent on using the visual-analog method of scaling in Experiments 3 and 4. People believed that learning would increase with more study repetitions both before and after completing the experiment.

The mean of participants' fitted slopes for their POLs, actual recall performance, pre-study global predictions, and post-study global predictions slopes are presented in the lower part of Table 1. All slopes were computed using participants' means on all four study-test trials except for POL slopes for groups who made predictions for only pairs studied once and four times. For this group, we could only compute POL slopes based on two data points (predictions for pairs studied once or four times). However, these slopes were equivalent to fitted POL slopes for people who predicted for all four study opportunities, $t < 1$, and hence we collapsed across this variable when computing means in Table 1. Prestudy global prediction slopes, $t(77) = 11.96$, $p < .001$, $d = 1.35$, POL slopes, $t(77) = 7.57$, $p < .001$, $d = .86$, post-study global prediction slopes, $t(77) = 10.83$, $p < .001$, $d = 1.23$, and recall slopes, $t(77) = 8.83$, $p < .001$, $d = 1.70$, were all significantly different from zero.

Consistent with previous findings, POL slopes were reliably related to pre-study global prediction slopes. Both POL slopes and pre-study global predictions slopes were related to actual recall performance (Table 2). However, unlike Experiment 3, the regression coefficient relating POL slopes to recall slopes ($\beta = .13$, $p = .29$) was not significant when controlling for the significant effect of pre-study global prediction slopes ($\beta = .23$, $p = .07$); $R^2 = .09$, $adjusted\ R^2 = .07$, $F(2, 75) = 3.78$, $p < .05$.

Post-study global predictions were most strongly related to recall slopes (Table 2). A model regressing post-study global prediction slopes on pre-study global prediction slopes, POL slopes, and recall slopes accounted for almost half of the variance in post-study global prediction slopes, $R^2 = .46$, $adjusted\ R^2 = .44$, $F(3, 74) = 21.24$, $p < .001$. The regression coefficient for recall slopes affecting post-study global predictions slopes ($\beta = .11$, $p = .26$) was significant even when controlling on pre-study global prediction slopes ($\beta = .11$, $p = .24$) and POL slopes ($\beta = .60$, $p < .001$). These results are consistent with previous evidence supporting the conclusion that people change their expectations about future learning based on monitoring their current task performance.

In summary, the current experiment revealed that people do not always overvalue stimulus features such as font size when predicting later memory. Although people's predictions were influenced by font size, they were influenced even more by how much study an item would receive. The effects of study on these predictions were not due lack of spatial reminders that study would occur causing font size cues to overshadow study cues or due to the contrast between the number of study repetitions for which people made their predictions.

Concerning the joint influence of font size and beliefs about study on POLs in the current experiment, the source of font size's effect on metacognitive judgments has recently come into question. Rhodes and Castel (2008) initially argued that the font size effect is caused by perceptual fluency. However, Mueller, Dunlosky, Tauber, and Rhodes (2014) proposed that the font size effect is due to belief-based processing. They demonstrated that people believe words printed in a larger font are more memorable than words presented in a smaller font (though without directly linking item-level judgments to these beliefs) and they found no evidence that fluency mediates the effects of font size on metacognitive judgments. If the font size effect in the current experiment is due to a belief-based process, it would suggest that multiple beliefs about memory may exert a simultaneous influence on people's metacognitive judgments. In any case, what is clear is that people sample and incorporate multiple sources of evidence when constructing their metacognitive judgments (Ariel & Dunlosky, 2011; Hertzog, Hines, & Touron, 2013; Tauber & Rhodes, 2012b). How access to different evidence exerts a simultaneous influence on monitoring is an intriguing question that requires further attention.

General Discussion

In the current experiments, people under-predicted the amount of learning that would occur across study-test trials more when predictions were framed to emphasize the test trial as opposed to the study trial. These results indicated that the stability bias observed in several experiments (e.g., Kornell & Bjork, 2009) is partially due to framing POLs to emphasize testing over studying. People may not hold accurate beliefs about the contribution of testing to improved memory performance (Karpicke & Roediger, 2008; Kornell & Son, 2009), and hence, framing predictions to emphasize test repetitions may lead people to discount beliefs about learning when making POLs, even when people are told they will have multiple study trials to learn items. The evidence suggests that people do access beliefs about learning, even with test framing (Experiment 4), yet their POLs show shallower slopes than the expected learning manifested in global predictions about performance, which often aligned well with actual recall slopes.

The current results highlight the dangers of inferring the influence of beliefs (or lack of beliefs' influence) on metacognitive judgments without directly measuring these beliefs and empirically evaluating their influence on monitoring. Consider the mean aggregate predictions for POLs and for the global judgments elicited before and after Experiment 3 (Figure 5). The shallow predicted learning curve for item-level POLs might lead one to conclude that beliefs about learning improvements had a minimal influence on item-level predictions. This same conclusion might also be reached by evaluating the differences between the aggregate values of global predictions and item-level POLs. However, actually

evaluating the relationship between individuals' slopes for POLs and global predictions led to completely different conclusions. The predicted rate of learning for pre-study global judgments – which measured pre-experimental beliefs about multi-trial study benefits – were strongly associated with item-level POL slopes (see Table 2). The stability bias appears to be a divergence of mean trends that belies the connection between pre-experimental beliefs and individual differences in POL slopes. Thus, evaluating only the mean patterns in POLs would have failed to reveal important relationships between beliefs and predictions about future learning.

Taken together, the results of the current experiments argue that any stability bias observed with POLs does not originate from people holding inaccurate beliefs about rates of learning. Instead, people have relatively accurate perceptions about multiple study-test trial learning, but they discount these beliefs as cues when making POLs (as with test framing in Experiments 2 and 4).

Why do people discount their beliefs that study will influence memory under test framing? According to regulatory focus theory (Higgins, 1997), framing influences the degree to which people focus on gains or losses during self-regulatory processing. Test framing may focus learners more on losses than study framing because thinking about testing may orient learners to think about memory failures. This loss focus could cause learners to discount potential gains from studying that they expect to occur because the possibility of forgetting is highly salient. An alternative explanation for the effects of framing is that some participants in the test frame groups with low working memory spans may have forgotten that restudy would occur on future study-test blocks. The study frame could serve as a reminder of this important task dimension that test framing does not provide. If people did not know that multiple study trials would occur either because of memory failures or inadequate task knowledge from failing to sufficiently read the task instructions, then learning beliefs may be less accessible during test framed POLs. Although this outcome is possible, the results of Experiment 4 suggests that learning beliefs are accessible even for test framed POLs which would not be expected if participants lacked knowledge that study would occur.

Clearly, beliefs about learning are available to people during multi-trial learning tasks. What we do not know from the present experiments is whether people access their beliefs about learning when regulating their study during multi-trial learning. A failure to access relevant learning beliefs when making POLs could foreshadow a similar failure to access said beliefs when making study decisions (e.g., whether to restudy items). On the other hand, any stability bias manifested in POLs could exert little or no influence on people's study behavior because deciding whether to study an item may in itself activate beliefs about how study affects memory (the decision is by default study-framed). Although no experiments have explicitly evaluated the activation of learning beliefs during restudy decision, people often choose in actual learning contexts to engage in elaborative or maintenance rehearsal to enhance learning (e.g., Hessen, 2011) and they do selectively choose to restudy material. Such findings suggest a better understanding of the benefits of restudy and rehearsal to ensure learning and long-term retention (Pyc & Rawson, 2011) than might be expected if a stability bias in POLs reflected deficient beliefs about how to regulate learning.

Of course the example above concerns a situation that is likely to be study framed. Learners' metacognitive monitoring may sometimes be initiated by test framed prompts which could promote assessments of learning that are more susceptible to a stability bias. This may be particularly likely for students who are performance oriented (Pintrich, 2000) because their goals are often to earn a certain grade on an upcoming examination, and hence, they may be oriented to think how likely am I to remember this material on the next test (test framed) opposed to will my memory benefit from studying this material again (study framed)? It is unclear whether study framing or test framing is more representative of real world learning contexts. We suspect that real world metacognitive monitoring may be triggered by study framed information in some contexts and test framed information in others depending on learners' goals and the current task conditions during study.

Although people expected repeated study to improve learning in the current experiments and beliefs exerted a strong influence on POLs, people still under predicted the amount of actual performance improvement in every experiment except for Experiment 1. These findings indicate that the stability bias can persist even when beliefs are accessed and applied to form metacognitive judgments with study framing. This persistent underconfidence could be due to a number of factors. One explanation is that the stability bias reflects an anchoring-and-adjustment effect (Epley & Gilovich, 2006; Tversky & Kahneman, 1974). According to anchoring-and-adjustment accounts of metamemory judgments, people's judgments are drawn towards a psychologically meaningful anchor, such as the midpoint of a range of possible performance (Connor, Dunlosky, & Hertzog, 1997; Scheck, Meeter, & Nelson, 2004). If this anchor is below actual levels of performance, people's judgments will underestimate performance. If this anchor is above actual levels of performance, their judgments will overestimate performance.

Consider the underconfidence-with-practice effect in which people underestimate their performance more on later study-test trials than on earlier study-test trials. This effect has been argued to occur because memory performance improves across study-test trials to levels above the judgment anchor (England & Serra, 2013; Scheck & Nelson, 2005). The stability bias may be influenced by a similar anchoring-and-adjustment mechanism. People likely adjust their POLs upward from an anchor when informed that an item will receive multiple study opportunities. The amount of adjustment is in part influenced by their beliefs about the effects of study on memory. However, people may fail to adjust far enough away from the anchor to account for actual improvements in performance that occur.

A second potential explanation is that the under-estimation that characterizes the stability bias results from methods of scaling POLs by respondents. It is typical to assume that POLs and other metacognitive judgments are direct outcomes of subjective probability of recall, justifying comparison of discrepancies between judgments and performance as if they validly scale over- or under-confidence. This assumption may not be justified (see Keren, 1991; Erev, Wallsten, & Budescu, 1994; Gu & Wallsten, 2001). Hanczakowski, Zawadzka, Pasek, & Higham (2013) recently provided experimental evidence that the underconfidence-with-practice effect for JOLs is a scaling artifact that does not reflect true underconfidence in memory. They contrasted JOLs made with a 0-100% rating scale with JOLs framed as binary (yes/ no) wagers on future recall. The wagers did not manifest underconfidence,

undermining the argument that the scaled JOLs are a simple manifestation of subjective probability of recall. The same concern may apply to POLs even with visual analog scaling of responses. If so, this would not undermine our claim that test framing causes people to discount learning because POL slopes were always shallower for test framed POLs than study framed POLs.

Our results reinforce the idea that an adequate test of a hypothesis about influences of beliefs on metacognition requires explicit measurement of the relevant beliefs and their relationship to metacognitive judgments and behavior. Typical approaches to examining the influence of beliefs on metacognitive judgments often do not explicitly measure beliefs, but instead evaluate them in an indirect fashion. One method involves having people make judgments for individual items immediately before studying them (termed pre-study JOLs; see Ariel & Dunlosky, 2011; Castel, 2008; Mueller, Dunlosky, Tauber, & Rhodes, 2014; Mueller, Tauber, & Dunlosky, 2013). Another method is the observer-judge paradigm, which attempts to eliminate access to internally generated covert cues that people experience during study (often termed experiential cues) that could influence judgments (Brennan & Williams, 1995; Jameson, Nelson, Leonesio, & Narens, 1993; Matvey, Dunlosky, & Guttentag, 2001; Vesonder & Voss, 1985). Both methods eliminate the experience of studying items and hence people can only apply their beliefs about how the cues available to them during judgments influence memory.

The advantage of directly requesting judgments that capture the beliefs is that empirically measuring them, as in this study, enables using them to predict other metacognitive judgments and learning behavior. Having participants make global judgments without engaging in study of material (as in this study; see also Kornell et al., 2011, Experiment 3) is a useful way of measuring beliefs. However, unless and until one directly links these beliefs to people's item level judgments, one should not assume that measured beliefs mediate the relationship between a given cue and a specific metacognitive judgment.

One potential concern about our approach to measuring study beliefs is that our prestudy global predictions may be constructed at the time of the judgment and not reflect beliefs about the effects of study on memory (for similar arguments about responses to memory beliefs questionnaires, see Cavanaugh, Feldman, & Hertzog, 1998; Schwarz & Knauper, 1999). People's judgments about memory can be manipulated by procedures that influence accessibility of relevant information (e.g., Winkielman, Schwarz, & Belli, 1998). It is possible then that people made prestudy global predictions in the current experiments by thinking about the example word pair provided to them in the task instructions, and then did so again when generating POLs during the first study-test trial. If so, the relationship between prestudy global predictions and POLs we observed may be derived from a similar generation process for the judgments themselves and not from accessing beliefs about study benefits. However, multiple lines of evidence suggest that this hypothesis does not account for the data.

First, if prestudy global predictions and POLs are similar judgments that involve similar constructive processes that do not tap beliefs about memory, it is unclear why prestudy global prediction slopes were much steeper than POL slopes in our experiments. Second, an

abundance of evidence indicates that people's predictions about changes in personal attributes are formed by applying implicit beliefs about general stability and change (for review, see Ross, 1989). Third, people appear to form beliefs about how skills change with practice at a young age (e.g., Dweck & Leggett, 1988). Fourth, people's willingness to study material multiple times in lab tasks and in preparation for examinations in school suggests that they expect repeated study to improve memory performance (Hartwig & Dunlosky, 2012; Karpicke, Roediger, & Butler, 2009; Kornell & Bjork, 2009; Wissman, Rawson, & Pyc, 2012). Fifth, global predictions correlate significantly with responses to memory belief questionnaires (Hertzog, Dixon, & Hulltsch, 1990). Given all this evidence, it is highly unlikely that people do not have beliefs about how memory changes with repeated study that would inform their prestudy global predictions.

In summary, the current experiments conclusively show that people believe that multiple study repetitions will improve learning. These beliefs influence POLs even when those judgments display a stability bias, indicating that the beliefs were accessed but discounted when constructing the POL. The stability bias is malleable, depends on how judgments are framed, and does not imply that individuals believe that recall performance does not change over study trials.

Acknowledgments

This research was supported in part by a Ruth L. Kirschstein National Research Service Award (NRSA) Institutional Research Training Grant from the National Institutes of Health (National Institute on Aging), Grant #5T32AG000175. We thank Anmol Chhabria, Zurain Hassan, Daniel Litowitz, Robert Reagin, Aly Skulskaya, Gemariah Valencia, Cambre Winters, Catherine Wong, and John Zimmermann for assistance with data collection.

References

- Abelson, RP. Statistics as principled argument. Lawrence Erlbaum Associates; Hillsdale, NJ: 1995.
- Ariel R, Dunlosky J. The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition*. 2011; 39:171–184. [PubMed: 21264621]
- Bellamy N, Campbell J, Syrotuik J. Comparative study of self-rating pain scales in osteoarthritis patients. *Current Medical Research & Opinion*. 1999; 15:113–119. [PubMed: 10494494]
- Borkowski J, Carr M, Pressley M. "Spontaneous" strategy use: Perspectives from metacognitive theory. *Intelligence*. 1987; 11:61–75.
- Brennan SE, Williams M. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*. 1995; 34:383–398.
- Brunier G, Graydon J. A comparison of two methods of measuring fatigue in patients on chronic haemodialysis: Visual analogue vs. Likert scale. *International Journal of Nursing Studies*. 1996; 33:338–348. [PubMed: 8736478]
- Castel AD. Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*. 2008; 36:429–437. [PubMed: 18426071]
- Carroll M, Nelson TO. Effect of overlearning on the feeling of knowing is more detectable in within-subject than in between-subject designs. *The American journal of psychology*. 1993:227–235. [PubMed: 8338189]
- Cavanaugh JC, Feldman J, Hertzog C. Memory beliefs as social cognition: A reconceptualization of what memory questionnaires assess. *Review of General Psychology*. 1998; 2:48–65.
- Connor LT, Dunlosky J, Hertzog C. Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging*. 1997; 12:50–71. [PubMed: 9100268]

- Dunlosky J, Ariel R, Ross B. Self-regulated learning and the allocation of study time. *Psychology of Learning and Motivation*. 2011; 54:103–140.
- Dunlosky J, Hertzog C. Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging*. 2000; 15:462–474. [PubMed: 11014709]
- Dweck CS, Leggett EL. A social-cognitive approach to motivation and personality. *Psychological Review*. 1988; 95:256–273.
- England BD, Serra MJ. The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review*. 2012; 19:715–722. [PubMed: 22477335]
- Epley N, Gilovich T. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*. 2006; 17:311–318. [PubMed: 16623688]
- Erev I, Wallstern TS, Budescu D. Simultaneous overconfidence and underconfidence: The role of error in judgment processes. *Psychological Review*. 1994; 101:519–527.
- Finn B. Framing effects on metacognitive monitoring and control. *Memory & Cognition*. 2008; 36:813–821. [PubMed: 18604963]
- Gu HB, Wallsten TS. On setting response criteria for calibrated subjective probability estimates. *Journal of Mathematical Psychology*. 2001; 45:551–563. [PubMed: 11493014]
- Guilford, JP. *Psychometric theory*. McGraw-Hill; New York, NY: 1954.
- Halamish V, McGillivray S, Castel AD. Monitoring one's own forgetting in younger and older adults. *Psychology and aging*. 2011; 26:631–635. [PubMed: 21463057]
- Hanczakowski M, Zawadzka K, Pasek T, Higham PA. Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language*. 2013; 69:429–444.
- Hartwig MK, Dunlosky J. Study strategies of college students: are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*. 2012; 19:126–134.
- Hertzog, C. Repeated measures analysis in developmental research: What our ANOVA text didn't tell us. In: Cohen, SH.; Reese, HW., editors. *Life-span developmental psychology: Methodological contributions*. Erlbaum; New York: 1994. p. 187–222.
- Hertzog C, Dixon RA, Hultsch DF. Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging*. 1990; 5:215–227. [PubMed: 2378687]
- Hertzog C, Hines JC, Touron DR. Judgments of learning are influenced by multiple cues in addition to memory for past test accuracy. *Archives of Scientific Psychology*. 2013; 1:23–32.
- Hertzog C, Price J, Burpee A, Frentzel BJ, Feldstein S, Dunlosky J. Why do people show minimal knowledge updating with task experience: Inferential deficit or experimental artifact? *Quarterly Journal of Experimental Psychology*. 2009; 62:155–173.
- Hertzog C, Price J, Dunlosky J. How is knowledge generated about memory encoding strategy effectiveness? *Learning and Individual Differences*. 2008; 18:430–445. [PubMed: 19043596]
- Hessen E. Rehearsal Significantly Improves Immediate and Delayed Recall on the Rey Auditory Verbal Learning Test. *Applied Neuropsychology*. 2011; 18:263–268. [PubMed: 22074064]
- Higgins ET. Beyond pleasure and pain. *American psychologist*. 1997; 52:1280–1300. [PubMed: 9414606]
- Jameson A, Nelson TO, Leonesio RJ, Narens L. The feeling of another person's knowing. *Journal of Memory and Language*. 1993; 32:320–335.
- Karpicke JD, Butler AC, Roediger HL III. Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory*. 2009; 17:471–479. [PubMed: 19358016]
- Karpicke JD, Roediger HL 3rd. The critical importance of retrieval for learning. *Science*. 2008; 319:966–968. [PubMed: 18276894]
- Keren G. Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*. 1991; 77:217–273.
- Koriat A. Monitoring one's own knowledge during study. A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*. 1997; 126:349–370.

- Koriat A, Bjork RA, Sheffer L, Bar SK. Predicting one's own forgetting: the role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*. 2004; 133:643–656. [PubMed: 15584811]
- Kornell N, Bjork RA. A stability bias in human memory: overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*. 2009; 138:449–468. [PubMed: 19883130]
- Kornell N, Rhodes MG, Castel AD, Tauber SK. The ease-of-processing heuristic and the stability bias: dissociating memory, memory beliefs, and memory judgments. *Psychological Science*. 2011; 22:787–794. [PubMed: 21551341]
- Kornell N, Son LK. Learners' choices and beliefs about self-testing. *Memory*. 2009; 17:493–501. [PubMed: 19468957]
- Lineweaver TT, Hertzog C. Adults' efficacy and control beliefs regarding memory and aging: Separating general from personal beliefs. *Aging, Neuropsychology, and Cognition*. 1998; 5:264–296.
- Matvey G, Dunlosky J, Guttentag R. Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition*. 2001; 29:222–232. [PubMed: 11352205]
- Mueller ML, Dunlosky J, Tauber SK, Rhodes MG. The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*. 2014; 70:1–12.
- Mueller ML, Tauber SK, Dunlosky J. Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*. 2013; 20:37–384.
- Pintrich, PR. The role of goal orientation in self-regulated learning. In: Boekaerts, M.; Pintrich, PR.; Zeidner, M., editors. *Handbook of self-regulation*. Academic Press; NY: 2000. p. 451-502.
- Price PC, Yates JF. Judgmental overshadowing: Further evidence of cue interaction in contingency judgment. *Memory & Cognition*. 1993; 21:561–572. [PubMed: 8412709]
- Pyc MA, Rawson KA. Costs and benefits of dropout schedules of test–retest practice: Implications for student learning. *Applied Cognitive Psychology*. 2011; 25:87–95.
- Reips UD, Funke F. Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*. 2008; 40(3):699–704. [PubMed: 18697664]
- Rhodes MG, Castel AD. Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*. 2008; 137:615–625. [PubMed: 18999356]
- Roediger HL III, Karpicke JD. Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*. 2006; 17:249–255. [PubMed: 16507066]
- Ross M. Relation of implicit theories to the construction of personal histories. *Psychological Review*. 1989; 96:341–357.
- Scheck P, Meeter M, Nelson TO. Anchoring effects in absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language*. 2004; 51:71–79.
- Scheck P, Nelson TO. Lack of pervasiveness of the underconfidence-with-practice effect: boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General*. 2005; 134:124–128. [PubMed: 15702968]
- Serra MJ, England BD. Magnitude and accuracy differences between judgments of remembering and forgetting. *The Quarterly Journal of Experimental Psychology*. 2012; 65:2231–2257. [PubMed: 22630784]
- Shaughnessy JJ. Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning and Verbal Behavior*. 1981; 20(2):216–230.
- Schwartz, N.; Knauper, B. Cognition, aging, and self-reports. In: Park, D.; Schwarz, N., editors. *Cognitive aging: A primer*. Psychology Press; 1999. p. 233-252.
- Tauber SK, Rhodes MG. Measuring memory monitoring with Judgments of Retention Interval (JOR). *Quarterly Journal of Experimental Psychology*. 2012a; 65:1376–1396.
- Tauber SK, Rhodes MG. Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and aging*. 2012b; 27:474–484. [PubMed: 21942899]

- Tullis JG, Finley JR, Benjamin AS. Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*. 2013; 41:429–442. [PubMed: 23242770]
- Tversky A, Kahneman D. Judgments under uncertainty: Heuristics and Biases, *Science*. 1974; 185:1124–1131.
- Vesonder GT, Voss JF. On the ability to predict one's own responses while learning. *Journal of Memory and Language*. 1985; 24:363–376.
- Winkielman P, Schwarz N, Belli RF. The role of ease of retrieval and attribution in memory judgments: Judging your memory as worse despite recalling more events. *Psychological Science*. 1998; 9:124–126.
- Winne, PH.; Hadwin, AF.; Dunlosky, J.; Graesser, AC. Studying as self-regulated learning. In: Hacker, DJ., editor. *Metacognition in Educational Theory and Practice*. LEA; Hillsdale, NJ: 1998. p. 277-304.
- Wissman KT, Rawson KA, Pyc MA. How and when do students use flashcards? *Memory*. 2012; 20:568–579. [PubMed: 22671698]

Highlights

- Showed framing manipulates the stability bias in predictions of learning (POLs)
- Beliefs about learning were independently measured before and after the experiment
- Individual differences in pre-experimental learning beliefs influenced POL slopes
- POL slopes underestimated actual learning, implicating belief discounting
- Beliefs correlated more highly with actual learning after task experience

Figure 1.

Illustration of a prediction of learning (POL) trial for the study frame group in Experiment 1. For the current item, participants' prediction represents the subjective likelihood that they will remember the pair just studied on the third study test-trial (i.e., after 3 additional study opportunities).

Figure 2. Mean percentage of word pairs actually recalled and mean prediction of learning (POL) for each test for the test frame (top panel) and study frame (bottom panel) groups in Experiment 1. Error bars represent standard error of the mean.

Figure 3.

Mean percentage of word pairs actually recalled and mean prediction of learning (POL) for each test as a function of quantity frame for the test frame (top panel) and study frame (bottom panel) groups in Experiment 2. Error bars represent standard error of the mean.

Figure 4.

Illustration of pre-study and post-study global predictions with visual analog rating scales in Experiment 3. Participants could draw a learning curve by clicking each bar to indicate their predicted degree of learning after each study opportunity.

Figure 5.

Mean percentage of word pairs actually recalled, mean prediction of learning (POL), mean pre-study global prediction, and mean post-study global prediction for each test for the seperate prediction (top panel) and joint prediction (bottom panel) groups in Experiment 3. Error bars represent standard error of the mean.

Figure 6.

Mean percentage of word pairs actually recalled, mean prediction of learning (POL), mean pre-study global prediction, and mean post-study global prediction for each test for the test frame (top panel) and study frame (bottom panel) groups in Experiment 4. Error bars represent standard error of the mean.

Figure 7. Mean percentage of word pairs actually recalled and mean prediction of learning (POL) for each test (left panel) and mean pre-study global prediction and mean post-study global prediction for each test (right panel) in Experiment 5. Error bars represent standard error of the mean.

Table 1

Mean fitted slopes for predictions and actual recall across study-test trials in Experiment 3 and 4 with visual analog rating scales and Experiment 5 with numerical scales.

	Experiment 3		Experiment 4		Experiment 5	
	Study Frame	Test Frame	Study Frame	Test Frame	Study Frame	Test Frame
Pre-study Global Prediction	11.49 (.84)	—	10.59 (1.12)	7.59 (1.86)	8.46 (.71)	—
POL	5.75 (.52)	—	3.71 (1.02)	.88 (.70)	4.14 (.55)	—
Actual Recall	13.57 (.74)	—	13.39 (1.28)	11.44 (1.16)	9.63 (.64)	—
Post-study Global Prediction	10.67 (1.03)	—	10.79 (1.54)	9.57 (1.02)	6.45 (.60)	—

Note. Standard errors of the means are in parentheses. All slopes differ significantly from zero at $p < .05$ except for the POL slope for the test frame group in Experiment 4. There were no test frame groups in Experiment 3 and 5.

Table 2

Correlations between individuals' fitted slopes in Experiment 3 to 5.

Slope	Study Frame				Test Frame			
	1	2	3	4	1	2	3	4
Experiment 3								
1. Pre-study Global Prediction	—				—			
2. POL	.63*	—			—	—		
3. Actual Recall	.23	.37*	—		—	—	—	
4. Post-study Global Prediction	.45*	.35*	.59*	—	—	—	—	—
Experiment 4								
1. Pre-study Global Prediction	—				—			
2. POL	.42*	—			.35	—		
3. Actual Recall	.38*	.29	—		.15	.07	—	
4. Post-study Global Prediction	.42*	.28	.67*	—	.21	.16	.82*	—
Experiment 5								
1. Pre-study Global Prediction	—				—			
2. POL	.43*	—			—	—		
3. Actual Recall	.28*	.23	—		—	—	—	
4. Post-study Global Prediction	.33*	.30*	.66*	—	—	—	—	—

Note.

* Denotes the correlation is significantly different from 0 at $p < .05$.